scientific reports

Check for updates

OPEN

MEAST: a novel framework for analyzing gene set stability in single cell transcriptomics

Fadi Abou Choucha✉ & Claude Pasquier

Biological processes rely on the coordinated activity of multiple genes, forming biological pathways that regulate specific cellular functions. Gene set activity analysis tools are essential for understanding how groups of genes regulate these processes. However, accurately scoring the activity of the most effective genes within a gene set remains challenging. Here, we introduce MEAST (Modular Expression Analysis and Scoring Toolkit), a robust computational framework that combines truncated singular value decomposition for gene set scoring with a genetic algorithm to identify stable, high-contributing subsets of genes. We validated our approach using both simulated and real single-cell RNA-seq datasets. In real single-cell blood RNA-seq datasets, the activity scoring function accurately distinguished between cell subtypes. Moreover, when tested on bulk RNA data, MEAST successfully distinguished between lung adenocarcinoma and lung squamous cell carcinoma by accurately scoring marker gene activities. The main feature of MEAST is the stability assessment, which based on activity scoring and uses a genetic algorithm to identify the core active genes within a gene set. Our results showed that MEAST effectively eliminates all low-activity genes in single-cell RNA-seq data. This study presents an advanced framework to analyze scRNA-seq data in wide system biology applications including biomarker identification, disease subtype classification, drug response prediction, and pathway analysis.

Keywords Gene set analysis, Transcriptomics, Pathway activity, Singular value decomposition, Genetic algorithm, Biomarker identification

Gene set analysis methods are widely used for studying gene expression patterns in RNA sequencing data. These tools help evaluate the behavior of gene sets that play a role in specific biological processes, signaling pathways, or regulatory mechanisms[1,2]. Defining active gene sets associated with specific biological processes or phenotypes helps to describe underlying disease mechanisms. This strategy also aids in identifying potential therapeutic targets and cellular responses to specific stimuli[3,4].

Genes do not function in isolation; biological processes often involve complex and coordinated networks of genes[5,6], where coexpressed genes control a variety of biological processes[7]. Genetic signatures are gene sets regulated under biological development[8], associated with cell types[9], or with different diseases[10]. Gene sets can be identified from high-throughput RNA sequencing (RNA-seq) data using various methods.

After gene set identification, a common step is gene set scoring or enrichment[8,11]. By evaluating gene set activity, we can infer the functional states of cancer cells[12]. The activity evaluation of regulatory networks is used to understand how regulatory elements drive biological processes[13,14]. Moreover, gene set scoring helps understand pathway-level perturbations, which are key to uncovering the biological mechanisms underlying diseases[15]. Existing computational tools, including GSVA, ssGSEA, AUCell, PAGODA, and PROGENy, are widely used to score gene set activities in scRNA-seq data. GSVA computes enrichment scores directly from gene expression data. This method allows pathway-level analyses without differential expression testing[15]. The ssGSEA independently computes enrichment scores for each sample and gene set[16], quantifying activities at the single-cell level. AUCell evaluates gene set activity by calculating the area under the recovery curve for gene expression rankings[17]. PAGODA is used to characterize transcriptional heterogeneity by identifying overdispersed gene sets[18]. In contrast, PROGENy is designed to infer pathway activity based on the expression levels of downstream target genes[19]. Generally, rank-based approaches, such as GSVA, ssGSEA, and AUCell, cannot effectively handle covariance in gene expression profiles. In addition, these tools do not account for variability within gene sets. Instead, focusing on the most active genes, rather than including all genes, can facilitate downstream analyses[2,20].

Université Côte d'Azur, CNRS, I3S, Sophia-Antipolis, France. ✉email: fchoucha@hotmail.com

Therefore, evaluating gene set stability can help identify subsets of the original gene set with higher expression activity.

Indeed, biological and technical variation, along with the gene set identification method, can introduce unrelated genes into a gene set[21]. Recent studies have shown the importance of assessing gene set stability to prioritize the most consistent contributors to pathway activity. For instance, focusing on core active genes has significantly enhanced biomarker robustness in cancer datasets[10]. By refining gene sets to include only the most stable and active members, researchers can minimize noise and improve reproducibility and biological significance[22]. Yang[23] suggested that only a subset of genes from an enriched gene set is responsible for the associated phenotype. Lippmann[24] proposed a computational method to reduce large gene sets by ranking genes based on their importance within biological process hierarchies. This approach retains over 70% of the original set's functionality while reducing the number of genes[24]. However, the method is limited by its dependence on existing knowledge bases and the inherent biases in gene selection. Further research introduced the Linear Combination Test for Gene Set Reduction (LCT-GSR), a computational tool to identify core subsets of gene sets associated with continuous phenotypes[20]. The LCT-GSR method utilizes a sequential gene removal strategy based on the Significance Analysis of Microarrays (SAM) statistic, which assumes additive and independent contributions of individual genes. This approach might overlook synergistic gene interactions within a pathway.

Therefore, combining gene set scoring with stability assessment enhances gene set analysis by focusing on the most active genes within a predefined set. In this context, we present MEAST, a novel package designed to evaluate gene set activity and stability in scRNA-seq data. MEAST utilizes eigengene weights obtained through truncated singular value decomposition (Truncated SVD) on the expression matrix and genetic algorithm to assess the gene set stability.

This paper describes the methodology behind MEAST and demonstrates its application to RNA-seq data analysis, including bulk RNA, with focus on single-cell transcriptomics. Through several case studies, we show how to identify the most stable gene subsets from predefined gene sets using a genetic algorithm while evaluating gene set activity with the Truncated SVD scoring method in RNA expression data.

Materials and methods

Overview of gene set activity analysis using MEAST

MEAST requires an expression matrix and a gene set as inputs, working with both single-cell and bulk data. It calculates gene module eigengenes using truncated SVD and projects these weights into the expression matrix to create an activity matrix (Fig. 1A). The activity-scoring process includes data preprocessing, subsetting the expression matrix to include only genes in the set, calculating module eigengenes, computing activity scores, and optionally permutation test. The framework provides visualization tools for gene set activity across cells or cell groups. For each gene set, MEAST derives an eigengene from the subset matrix and projects its weights to generate activity values (Fig. 1A). By aggregating these values within specific groups, users measure gene set activity. Parameters like component number, scaling approaches, and aggregation methods can be adjusted to fit research needs.

For stability evaluation, MEAST uses genetic algorithms (GAs) to identify stable subsets within larger gene sets. GAs optimize gene sets based on activity in specific cell groups by selecting subsets that contribute to activity signals. The process starts with generating an initial population of gene subsets from a larger set. Each subset is evaluated for its activity score within a target cell group. Parents are selected based on fitness, then crossover and mutation create new gene combinations. The final output is the most stable and active gene subset from the original set G within the cell group (Fig. 1B).

Single cell RNA-seq dataset

The study utilized a pre-normalized read count matrix from[25], available on the Single Cell Portal[26]. Villani[25] conducted quality control and normalization through TPM normalization. The dataset comprised 1078 dendritic cells (DCs) and monocytes, with a median of 5326 genes per cell and 26,593 genes. The genetic markers examined matched those outlined in the original publication.

We also used a public Drosophila melanogaster ovary scRNA-seq dataset from ArrayExpress (E-GEOD-141701)[27]. In our analysis, we used the provided normalized expression matrix files from E-GEOD-141701. The dataset contains 1270 single cells from ovary tissue (mixed pool of 200 ovaries; wild-type genotype; normal). The main inferred cell-type groups are: follicle cell (468 cells), anterior escort cell (404 cells), posterior escort cell (388 cells), and germarium cap cell (6 cells). In the benchmarking experiments, we used follicle cells as the target population and compared them to all other annotated cells in this dataset.

Bulk RNA-seq data processing and normalization

RNA-seq read count data for Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) samples were obtained from TCGA through the GDC portal. As of July 2024, 833 LUAD and 708 LUSC samples were downloaded. Only primary tumor or normal tissue samples were included. Data were processed and normalized using DESeq2 (v1.30.1) in R (v4.0.3) with default settings. Raw read counts were extracted from the integrated counts matrix. Genes with zero counts across all samples were excluded. A prefiltering step removed low-expression genes (fewer than 10 counts across all samples).

Simulation of single-cell RNA data

scRNA-seq dataset was generated using SPARSim[28] with preset parameters from[29], modeled on the Zheng C1 cell type. Two non-overlapping gene sets (Set A and Set B) of 200 genes each were defined to evaluate gene activity scoring. Set A fold changes were sampled uniformly between 5 and 25, then scaled in a dose-dependent manner: ×1, ×2, ×3, and ×4 for Doses 1–4, simulating proportional increases in gene expression with treatment

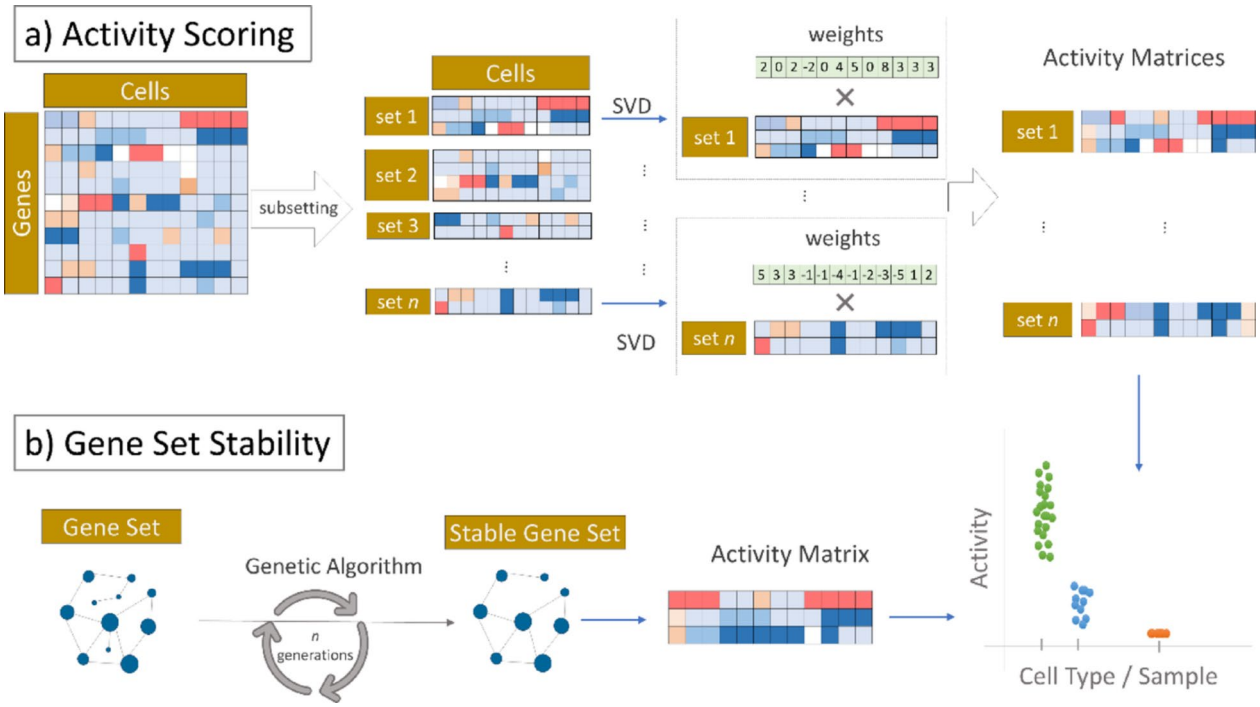

Fig. 1. Overview of MEAST: gene set activity scoring and stability analysis. (**a**) Gene Set Activity Scoring. MEAST quantifies gene set activity using singular value decomposition (SVD) on single-cell or bulk expression data. The input expression matrix (left) contains gene expression values across multiple cells (or samples). Gene set-specific expression matrices are extracted as subsets of the original matrix. SVD is applied to these subsets to derive eigengenes, whose corresponding weights are then projected back onto the original gene set-specific matrices, generating an activity matrix. This process quantifies the collective activity of each gene set across cells or samples. The resulting activity scores allow for downstream comparisons across different cell types or conditions (right), providing insights into gene module regulation in distinct biological states. (**b**) Gene Set Stability Analysis. MEAST integrates a genetic algorithm (GA)-based optimization framework to refine gene sets for stability and functional relevance. The GA iteratively optimizes gene subsets by assessing their activity consistency across specific cell types or conditions. The process begins with an initial gene set (left), which undergoes selection, crossover, and mutation across multiple generations to identify a stable subset with the highest activity signal (middle). This optimized subset is then used to compute an activity matrix. The package offers (right) demonstrates the gene set activity across cell types or samples, enabling enhanced biological interpretation.

intensity. Set B maintained high expression across all doses, with fold changes between 5 and 10, serving as a stable positive control to assess scoring method robustness and specificity.

Simulations included four treatment doses with cell counts of 750, 900, 700, and 1000 for Doses 1–4, along with 1000 untreated control cells representing baseline expression. Differential gene expression (DGE) parameters were generated using the *SPARSim_create_DE_genes_parameter* function, including transcriptional noise to simulate biological variability. The final dataset consisted of 4350 cells (3350 treated and 1000 control). Transcriptional noise was included to replicate variability observed in real scRNA-seq experiments.

Statistical analysis

To evaluate correlation between dose levels and gene set activity scores, Kendall's Tau correlation test was performed using the SciPy library's Kendalltau function. For Gene set A, analysis included all dose levels and the control group. For gene set B, correlation analysis was restricted to treatment doses. Statistical significance was determined using a p-value threshold of 0.05. Differences in activity scores were evaluated using two-sided Mann-Whitney U test to compare distributions between target clusters and other clusters. P-values were adjusted using the Benjamini-Hochberg procedure, with adjusted p-values below 0.05 considered statistically significant.

MEAST methods—activity scoring

The MEAST framework calculates gene set activity at individual cell and cell group levels through data preprocessing, gene set subsetting, Module Eigengene calculation, activity score computation, and permutation-based statistical significance assessment.

Data preprocessing and gene set subsetting

Gene sets, which represent genes related to specific biological processes or pathways, serve as the foundational units for activity analysis. For each gene set, the expression matrix is subsetted to include only the genes within

the set. This results in an expression matrix E_g with dimensions *G x M*, where *G* indicates the number of genes in the set, and *M* represents the number of cells.

Formally, given the entire gene expression matrix *E* with dimensions *N*×*M* (where *N* is the total number of genes) and a gene set *S* comprising *G* genes:

$$E_g = E[S, :]$$

Here, E_g is the subset of *E* containing only the genes in *S*.

Module Eigengene calculation

The module Eigengene is calculated using dimensionality reduction techniques. Optionally, gene expression data may be standardized to have a mean of zero and unit variance for normalization across genes.

$$E_G = StandardScaler\left(E_G^T\right)$$

where G_G^T denotes the transpose of E_G. *StandardScaler* is a sklearn function.

By default, MEAST uses Truncated SVD for dimensionality reduction. Alternatively, PCA may be used. The number of principal components retained is represented by *n* components, with a default value of 1, corresponding to the component explaining the most variance. The first principal component (PC1) is extracted as the Module Eigengene ω:

$$\omega = V_{(:,1)}$$

where *V* is the matrix obtained from SVD.

Alternatively, the absolute values of the eigenvector loadings can account for both positive and negative contributions:

$$\omega = \left|V_{(:,1)}\right|$$

Activity score calculation

The activity score quantifies the gene set's collective behavior within each cell or cell group by projecting gene expression data onto the Module Eigengene. For individual cells, the activity score vector *a* a is calculated as:

$$a = E_G^T \cdot \omega$$

where E_G^T is the transposed gene expression matrix (*M*×*G*), ω is the Module Eigengene (*G*×1), a is the activity score vector (*M*×1).

For predefined cell groups, the Group Activity Score $Activity_g$ for group *g* is computed as the mean of activity scores across all cells in the group:

$$Activity_g = \frac{1}{|C_g|} \sum_{c \in Cg} a_c$$

where C_g is the set of cells in group *g*, a_c *is the activity score of cell c.*

MEAST evaluates statistical significance through permutation testing, creating a null distribution by randomly shuffling cell labels, then recalculates activity scores after each permutation using the same Module Eigengene. This process repeats thousands of times for robust null distribution estimation.

MEAST methods—stability assessment

MEAST evaluates gene set stability within specific cell populations using a Genetic Algorithm (GA) that identifies key gene subsets with highest activity levels.

Stability analysis filters through different gene subsets S, retains only co-expressed genes with higher activity in specified cell populations. MEAST starts with a complete gene expression matrix *E* (rows=genes, columns=cells). From matrix *E*, a gene set *G* and target cell group C_g are defined to measure gene activity within the target population. In the stability function, user defines initial population size, determining how many primary subsets the GA algorithm generates. The GA fitness score is defined by gene subset activity, calculated by projecting expression data onto a Module Eigengene *w* obtained through Truncated SVD or PCA. This follows the same activity scoring method from the MEAST framework.

Mathematically, the activity score a_c for each cell $c \in C_g$ is calculated as:

$$a_c = S^T E[:, c] \cdot w$$

where $E[:, c]$ denotes the expression profile of cell c across the genes in subset S. The overall fitness of the gene subset is then determined by averaging these activity scores across all cells in C_g:

$$Fitness(S) = Activity_g(S) = \frac{1}{|C_g|} \sum_{c \in Cg} a_c$$

The GA algorithm selects parent subsets using tournament selection, which favors individuals with higher fitness while maintaining genetic diversity. Crossover operations (uniform or single-point) then recombine genetic material from parents to produce offspring with potentially superior characteristics. Mutation operations randomly alter elements of offspring's binary vectors based on a specified rate to enhance diversity and prevent premature convergence. This selection, crossover, and mutation process continues for a predetermined number of generations or until convergence criteria are met.

Cohen's d calculation

Cohen's d was calculated as a standardized effect size measure to quantify the discriminative power of each scoring method. For each gene set and method, Cohen's d measured the separation between activity scores in the target cell population versus all other cells. The metric was computed using pooled standard deviation as follows:

$$d = 1 + \frac{\bar{x}_{target} - \bar{x}_{non-target}}{s_{pooled}}$$

Higher Cohen's d values indicate stronger separation between target and non-target populations, reflecting better discriminative performance. The absolute value was reported to ensure positive values regardless of direction.

Robust coefficient of variation (rCV)

We calculated a robust coefficient of variation (rCV) using median absolute deviation (MAD) normalized by the median:

$$rCV = \frac{MAD}{median}$$

This metric complements Cohen's d by evaluating within-group homogeneity rather than between-group separation.

Results

Gene set activity scoring

Activity scoring with simulated scRNA-Seq data

To evaluate the gene set activity scoring method in MEAST, we analyzed a simulated single-cell RNA sequencing (scRNA-seq) dataset generated using SPARSim[28]. The dataset included two distinct gene sets: Gene Set A, designed for dose-dependent expression increases at four dose levels, and Gene Set B, designed to maintain consistently high expression levels across all four doses (Table S1) (see Materials and methods section).

We calculated activity scores for the two gene sets across all treatment doses and the control, followed by Z-score scaling. For Gene set A, the activity scores increased proportionally with dose levels, ranging from − 1.00 in the control group to − 0.87, − 0.34, 0.56, and 1.67 for Doses 1 through 4, respectively (Table S2). We used Kendall's Tau to analyze statistically the correlation between dose levels and activity, where a coefficient of 1.00 ($p = 0.017$) shows a strong positive correlation between augmenting dose and corresponding gene set activity. This agrees with the design of Gene set A, where higher doses were expected to cause proportional increases in activity (Fig. 2A, C). On the other hand, Gene set B scores showed stable activity across the four levels, with values of − 2.00 for the control group and 0.50, 0.47, 0.56, and 0.46 for Doses 1 through 4, respectively (Table S2). Excluding the control group, the Kendall's Tau score was − 0.33 ($p = 0.75$), indicating unchanged activity across dose levels, aligning with Gene set B's stable activity design (Fig. 2B, C).

Activity scoring with single cell RNA real data

Dendritic cells (DCs) in human blood have been classified into multiple subtypes, though their exact number and interrelationships remain uncertain. Recently, *the published study*[25] utilized scRNA-seq on approximately 1,200 high-quality single cells from a healthy donor. Their analysis identified six distinct DC subtypes and four monocyte subtypes using a total of 242 marker genes, all with an Area Under the Curve (AUC) of at least 0.85. The Figure S1 shows the sizes of these gene sets used in the current study. They demonstrated that the gene set markers: CD141/CLEC9A, CD1C-A, CD1C-B, CD1C–CD141–, AS DCs, and pDC were specific to the newly defined DC clusters: DC1, DC2, DC3, DC4, DC5, and DC6, respectively.

To further elucidate the molecular marker sets that distinguish these dendritic cell (DC) subtypes, we applied our activity scoring method to compute marker activity scores at both DC cells clusters (DC1, DC2, DC3, DC4, DC5, and DC6) and single-cell levels. We showed that the gene set CD141/CLEC9A exhibited a high activity score in DC1 (2.13), significantly higher than in other clusters (− 0.08 to − 0.46; Mann–Whitney U test, adjusted $p = 2.42 \times 10^{-92}$) (Table S3). The higher activity of CD141/CLEC9A differentiates DC1 cells from other subtypes (Fig. 3A) and confirms its specificity for the DC1 subtype. Similarly, the gene set CD1C-A has high activity in both DC2 (2.12) and DC3 (1.75), with significantly lower scores in other cell types (adjusted $p = 3.19 \times 10^{-50}$ for DC2; 2×10^{-37} for DC3) (Table S3). These suggest that CD1C-A serves as a shared marker for both DC2 and DC3 subtypes (Fig. 3B), albeit with stronger activity in DC2. In contrast, the CD1C-B gene set showed a high activity score in DC3 (0.98), significantly higher than in other clusters (adjusted $p = 2 \times 10^{-37}$) (Table S3), indicating that CD1C-B is a more unique and distinctive marker for the DC3 subtype (Fig. 3C). The finding aligns with the reference study by Villani[25], which identified CD1C as the sole marker uniquely shared by DC2 and DC3. Notably, our scoring method quantitatively shows that CD1C-B is more specific to DC3 than CD1C-A and enables precise activity quantification. The gene set CD1C–CD141– exhibited high activity in DC4 (mean = 2.22) and significantly lower activity in other clusters (adjusted $p = 1 \times 10^{-82}$), confirming it as a specific marker

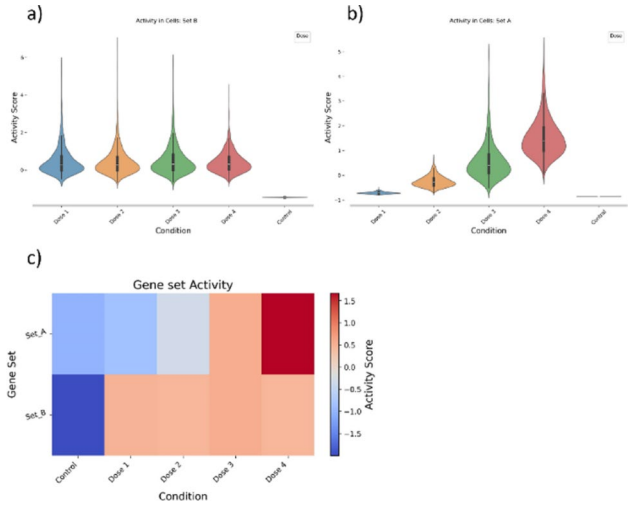

Fig. 2. Gene Set Activity Analysis Using MEAST on Simulated single-cell RNA-seq Data. Gene Sets A and B activity scores were calculated with MEAST across treatment doses and visualized to assess dose-dependent gene expression. (**a**) The violin plot of Gene Set A activity scores across treatment conditions shows a dose-dependent increase (Kendall's Tau = 1.00, $p = 0.0167$). (**b**) The violin plot of Gene Set B activity scores shows minimal variability across doses (Kendall's Tau = −0.33, $p = 0.75$). (**c**) Heatmap of Z-score scaled activity scores for Gene Sets A and B across control and four increasing dose levels. Gene Set A shows an evident dose-dependent upregulation, increasing activity scores as doses rise. Gene Set B remains stable, with minor fluctuations across doses.

for DC4 (Fig. 3D). The AS DCs gene set demonstrated significant activity in DC5 (1.99) but showed minimal activity in other clusters (Fig. 3E), supporting its specificity for DC5 (adjusted $p = 1 \times 10^{-20}$). Finally, the pDC gene set was predominantly active in DC6 (2.11), with negligible activity across DC1 to DC5 (− 0.30 to 0.38; adjusted $p = 1.6 \times 10^{-95}$), further confirming its specificity for DC6 (Fig. 3F). These findings support and extend the DC classification proposed by Villani[25] and provide an accurate way to score the activity and show the specificity of predefined markers.

Activity of biomarkers in LUAD and LUSC RNA-seq data

Lung cancer is a major cause of cancer deaths worldwide, with lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) as the primary subtypes. Accurate differentiation between LUAD and LUSC is essential for prognosis and treatment, yet remains challenging. Chen and Dhahbi[30] identified 17 potential discriminative genes, five specific to LUAD and twelve to LUSC.

In this study, we employed MEAST to assess the expression activity of these 17 marker genes using RNA-seq data from the Cancer Genome Atlas (TCGA). Expression analysis methods, such as those proposed by[30], may have limited effectiveness in distinguishing LUAD from LUSC, as these cancer types are closely related. We computed activity scores for LUAD and LUSC marker sets using an expression matrix of 14,010 genes from 833 LUAD and 708 LUSC bulk RNA-seq samples. Scores were calculated using the mean expression method and then normalized. Finally, we grouped the scores by sample origin into four categories: Normal LUAD, Normal LUSC, Tumor LUAD, and Tumor LUSC.

For the LUAD marker gene set, activity scores were significantly higher in Tumor LUAD samples (1.64) than in Tumor LUSC (−0.87), Normal LUAD (−0.76), and Normal LUSC (−0.02) (Mann–Whitney U test, adjusted $p < 1.00 \times 10^{-20}$) (Table S4; Fig. 4A). Similarly, the LUSC marker gene set had markedly higher scores in Tumor LUSC (mean = 1.73) compared to Tumor LUAD (−0.57), Normal LUAD (−0.58), and Normal LUSC (−0.58) (Mann–Whitney U test, adjusted $p < 1.00 \times 10^{-20}$) (Table S4; Fig. 4B). These strong differences confirm MEAST's ability to distinguish LUAD from LUSC using gene sets identified from bulk RNA-seq data (Fig. 4C).

Gene set stability assessment

Stability evaluation in simulated data

We applied the stability method to a simulated scRNA-seq dataset across two previously generated treatment conditions: Dose 2 and Dose 4. We used the following parameters: a population size of 50, 500 generations, a mutation rate of 0.001, and a crossover rate of 0.5.

In Dose 2, activity scores increased progressively over initial generations, reaching 4883 at generation 295 (Fig. 5A). The Dose 4 group showed a similar pattern, starting with a baseline activity score of 414 and achieving an 11.6-fold increase to 4811 after 171 generations (Fig. 5B). Both optimizations resulted in 6 genes while excluding 194 genes from the original pool (Fig. 5C, D).

To assess the robustness of the stability method, we added 6 randomly selected genes to previously optimized gene sets from Dose 2 and Dose 4 treatments, creating 12-gene sets with deliberate noise. We performed 5 independent GA runs using identical parameters. In all runs, the algorithm rapidly eliminated all noise genes

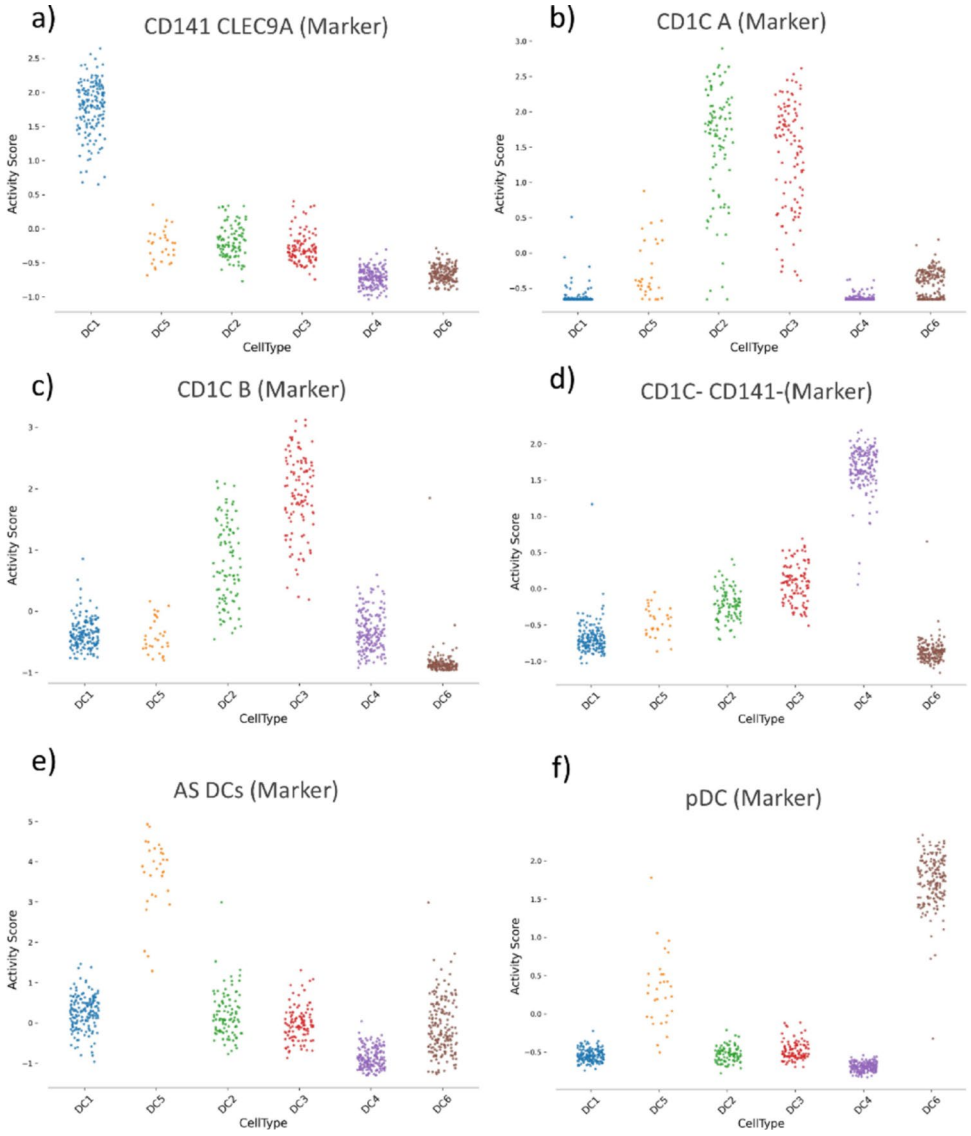

Fig. 3. Activity scores of marker gene sets in dendritic cell (DC) subtypes. Dot plots illustrate the activity scores of marker gene sets across six DC subtypes (DC1–DC6). (**a**) The AS DCs gene set is notably enriched in DC5. (**b**) The CD1C A gene set is highly active in both DC2 and DC3, serving as a common marker for these subtypes. (**c**) The CD1C_B gene set shows specificity for DC3. (**d**) The CD1C–CD141– gene set is distinctly enriched in DC4. (**e**) The CD141/CLEC9A gene set is predominantly enriched in DC1. (f) The pDC gene set displays exclusive enrichment in DC6. These results highlight the unique molecular signatures of DC subtypes and support a more refined classification based on gene activity profiles.

within the first 6 generations for Dose 2 (median = 6) and within the first 5 generations for Dose 4 (median = 5) (Table S5). Final fitness scores matched those from the original runs for both gene sets (Fig. 6).

Stability and gene set size reduction in real scRNA-seq data

To evaluate the MEAST algorithm's performance on real scRNA-seq data, we applied the stability to the same human blood cell dataset[25] previously used in the current study. For this purpose, we initially focused on optimizing the pDC gene set within the DC6 subtype and then processing all gene sets across DC subtypes and monocyte populations.

We evaluated the stability of the gene set pDC within the cell type DC6, where we had previously identified pDC as a specific marker for DC6. For that, we used the following parameters: a population size of 100, 1,000 generations, a mutation rate of 0.001, a crossover rate of 0.5, and a minimum gene set size of 50 genes. The best activity score of 320 was achieved after 507 generations (Fig. 7A), resulting in an optimized gene set of 99 genes while excluding 291 genes from the original pDC gene set (Fig. 7B). We then introduced 25 randomly selected genes into the optimized set to assess robustness, creating a noised gene set of 124 genes. The GA excluded this noise by the 6th generation, restoring the fitness score to 320 (Fig. S2). Analysis of the activity curve in Fig. 7A shows stabilization around the 450th generation, indicating 500 generations are sufficient for stable gene set

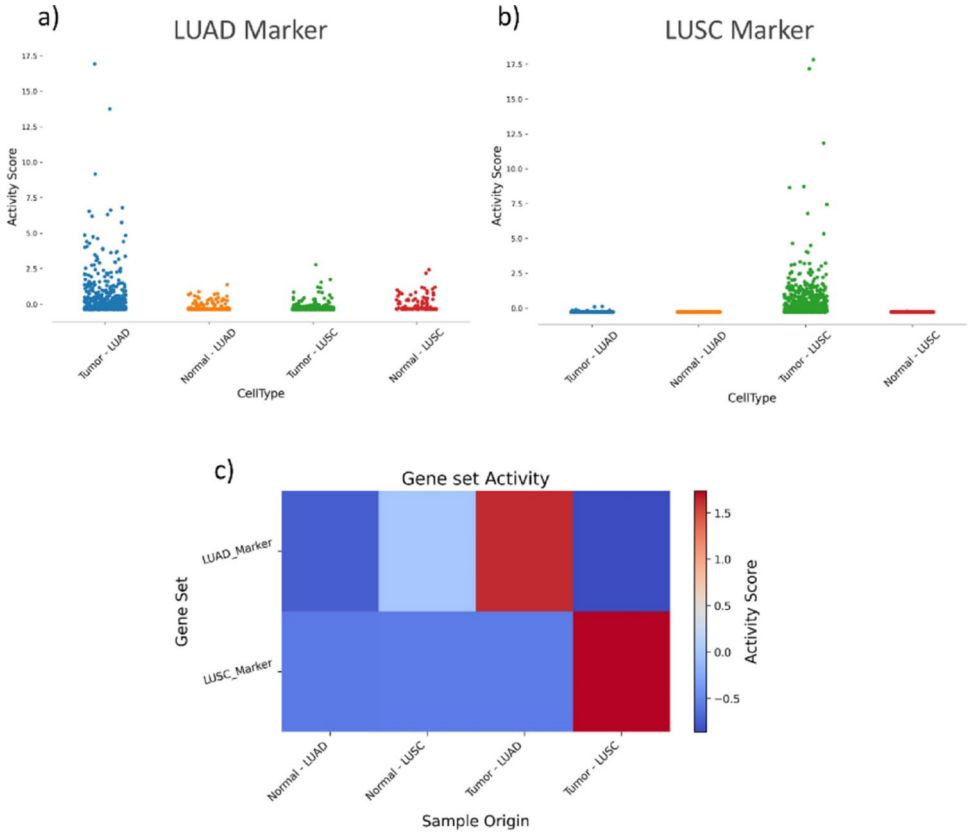

Fig. 4. Activity scores of LUAD and LUSC marker gene sets. Plots display MEAST-derived activity scores for LUAD (**a**) and LUSC (**b**) marker gene sets across four sample categories: Tumor LUAD, Normal LUAD, Tumor LUSC, and Normal LUSC. LUAD markers show higher activity in Tumor LUAD, while LUSC markers show increased activity in Tumor LUSC, confirming subtype-specific expression patterns. Each dot represents a sample. (**c**) The heatmap displays activity scores for LUAD and LUSC marker gene sets calculated by MEAST across various sample origins: Normal LUAD, Normal LUSC, Tumor LUAD, and Tumor LUSC. LUAD markers show the highest activity in Tumor LUAD, while LUSC markers display increased activity in Tumor LUSC, highlighting the distinct differentiation of subtypes. The color gradient illustrates normalized activity, with blue representing lower scores and red indicating higher scores.

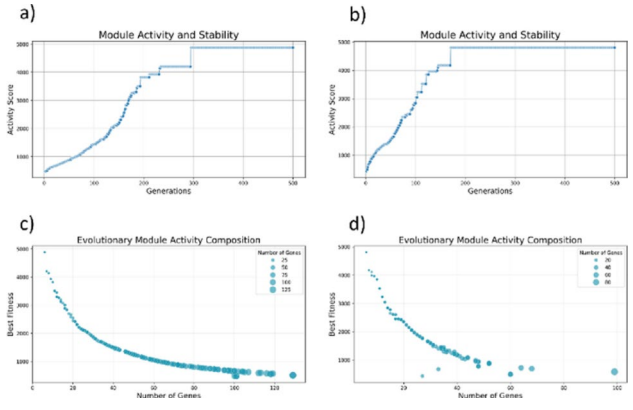

Fig. 5. Stability assessment of gene Set B under Dose 2 and Dose 4 treatments. (**a**, **b**) Optimization of gene set stability to maximize activity while focusing on key genes. Activity score progression over 500 generations for Dose 2 (a) and Dose 4 (**b**). (**c**, **d**) Relationship between gene count and fitness (activity) score for Dose 2 (**c**) and Dose 4 (d), showing an inverse correlation, indicating effective exclusion of non-contributory genes. The genetic algorithm (GA) optimized gene subsets from an initial set of 200 genes, increasing activity scores to 4883 for Dose 2 (**c**) and 4811 for Dose 4 (**d**), retaining 6 core active genes for both.

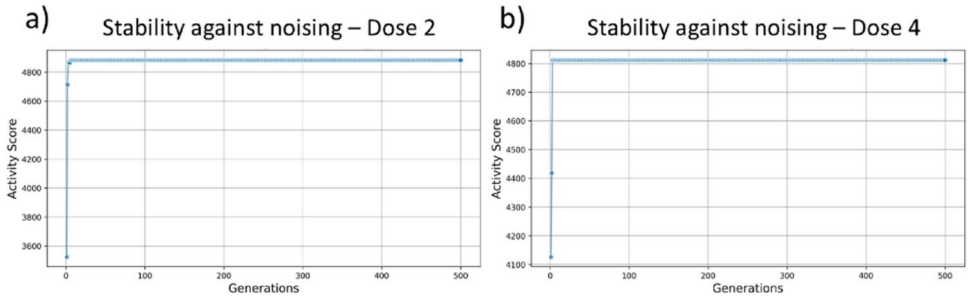

Fig. 6. Gene set refinement using the stability function. Results from the MEAST algorithm optimization for Dose 2 (**a**) and Dose 4 (**b**). The stability function eliminates noisy genes in early generations while keeping activity scores identical to their initial values. Activity scores were evaluated over 500 generations (X-axis) and plotted against activity score values (Y-axis).

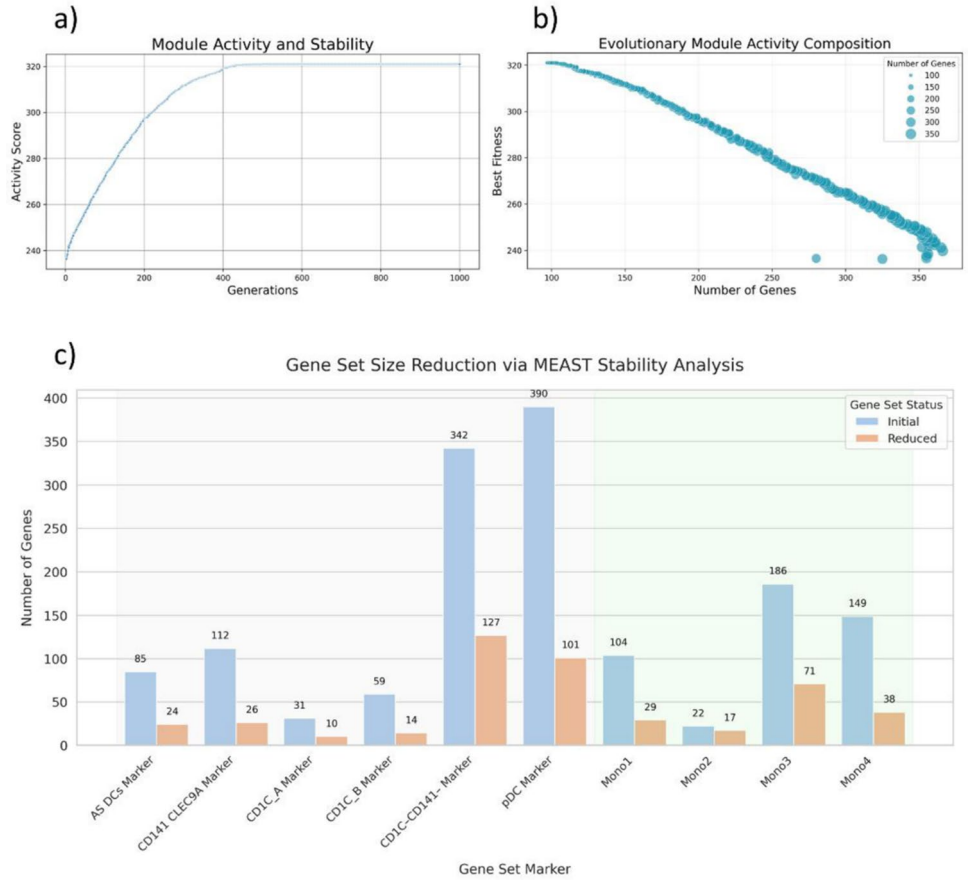

Fig. 7. Stability and Gene Set Size Reduction in Real scRNA-seq Data. (**a**) A scatter plot shows the optimization of activity scores over 1000 generations, with an increase from 240 to a peak of 320 at generation 507. (**b**) The scatter plot illustrates the inverse relationship between gene count and fitness score, indicating that larger gene sets correspond to lower fitness in pDC. (**c**) A bar chart depicts the reduction in gene set size across the dendritic cell (DC) and monocyte marker sets using the stability function in MEAST. On average, gene sets were reduced by 73% for DC markers and 65% for monocyte markers.

composition. Intersection analysis confirmed no introduced noise genes were retained in the final optimized gene set, while all original 99 genes were preserved.

Next, we intended to utilize the stability function to reduce the size of gene sets. We included all DC gene sets: CD141/CLEC9A, CD1C-A, CD1C-B, CD1C–CD141–, AS DCs, and pDC — and Mono-1 to -4. We configured the genetic algorithm with 500 generations, a population size of 100, a mutation rate of 0.001, a crossover rate of 0.5, and a minimum gene count of 10. The gene set AS DCs was reduced from 85 to 24 genes, the CD141 CLEC9A from 112 to 26 genes, the CD1C-A from 31 to 10 genes, the CD1C-B from 59 to 14 genes, the CD1C-

Metric	Unit	Range (min–max)	Mean ± SD
Execution time	Second	4–6	4 ± 0.6
CPU usage	Percent	56–92	87 ± 11
Peak memory	Megabyte	403–462	429 ± 18
Avg memory	Megabyte	279–364	330 ± 24

Table 1. Gene set activity analysis performance metrics. All measurements obtained over 10 runs using 40 CPU cores. MEAST gscore function applied to human blood cell dataset from Villani et al.[25].

Metric	Unit	Range (min–max)	Mean ± SD
Execution time	Second	635–653	645 ± 8
CPU usage	Percent	2875–2910	2883 ± 15
Peak memory	Megabyte	423–458	438 ± 17
Avg memory	Megabyte	411–454	428 ± 19

Table 2. Gene set stability analysis performance metrics. ll measurements obtained over 5 runs using 40 CPU cores. Parameters: Population size = 100, Generations = 1,000, Mutation rate = 0.001, Crossover rate = 0.5, Min gene set size = 50.

CD141—from 342 to 127 genes, and the pDC from 390 to 101 genes (Fig. 7C). The size of monocyte-specific gene sets was decreased as follows: Mono1 from 104 to 29, Mono2 from 22 to 17, Mono3 from 186 to 71, and Mono4 from 149 to 38. The MEAST stability function effectively reduced the initial gene sets to smaller, more active subsets, yielding higher activity scores.

Computational performance of MEAST in gene set activity and stability analyses

We assessed the MEAST package's computational efficiency on a 40-core high-performance computing cluster. Table 1 shows gene set activity analysis using the MEAST gscore function on the Villani[26] dataset has high efficiency (execution time: 4.2 ± 0.64s, CV = 15.2%) with moderate CPU use (87.27 ± 11%) and reasonable memory consumption (429 ± 18 MB). Table 2 shows stability analysis using the MEAST stability function (population size = 100, generations = 1,000, mutation rate = 0.001, crossover rate = 0.5, minimum gene set size = 50) requires more resources, with longer execution time (645 ± 8s) and higher CPU utilization, while memory requirements remained similar (438 ± 17 MB). Memory utilization showed better consistency (CV = 4.2–7.2%) than execution time in both analyses.

Benchmarking MEAST activity scoring and evaluation of stability function

MEAST comprises two integrated components: an activity scoring method based on truncated SVD, and a genetic algorithm GA-based stability assessment module that identifies core active genes. While the stability module represents the primary methodological innovation of MEAST, the activity scoring component can be directly compared with established methods. We therefore performed benchmarking on both the full gene sets and the GA-stable subsets.

We compared MEAST activity scoring against three widely used methods: GSVA, ssGSEA, and AUCell. To quantify discriminative power, we used Cohen's d as a separation metric, measuring the standardized difference between target and non-target score distributions. Cohen's d provides a scale-independent effect size and allows direct comparison across datasets and gene sets.

Using the human blood dendritic cell dataset from Villani[25], we computed activity scores for DC subtype marker gene sets and calculated Cohen's d between each target DC subtype and all other cells. Averaged across all marker gene sets, the mean absolute Cohen's d values were 3.548 for MEAST, 3.196 for AUCell, 2.784 for GSVA, and 2.719 for ssGSEA (Fig. 8A). In the subtype-level summary, MEAST gave the highest mean absolute Cohen's d for each of the four retained DC contrasts: DC1 (7.52 for MEAST vs. 6.08 for AUCell, 5.58 for ssGSEA, and 5.33 for GSVA), DC2 (2.28 vs. 2.18, 1.82, and 1.93), DC3 (3.58 vs. 3.06, 2.52, and 2.53), and DC4 (2.64 vs. 2.62, 2.13, and 2.31) (Fig. S4, supplementary data). These results indicate that MEAST activity scoring effectively distinguishes target cell populations from non-target cells.

To assess performance on an independent dataset, we analyzed a Drosophila ovary scRNA-seq dataset (E-GEOD-141701)[27] using three curated KEGG pathways (TGF beta signaling pathway, Wnt signaling pathway, and Insect hormone biosynthesis), with follicle cells as the target group. Mean Cohen's d across pathways (follicle cell vs. all other cells) showed: AUCell = 0.656, MEAST = 0.370, GSVA = 0.350, and ssGSEA = 0.252 (Fig. 8B). In the other cell types MEAST was highest for anterior escort cells (mean absolute Cohen's d = 0.723) and GSVA was highest for posterior escort cells (0.554) (Fig. S4, supplementary data).

To evaluate whether the GA stability module improves downstream analysis, we tested whether GA-selected gene subsets enhance both score consistency and discriminative power across all four scoring methods. Using the same Drosophila ovary dataset with the three KEGG pathways and follicle cells as the target, the GA reduced gene set sizes while improving within-target score stability (Fig. 8C): Insect hormone biosynthesis decreased

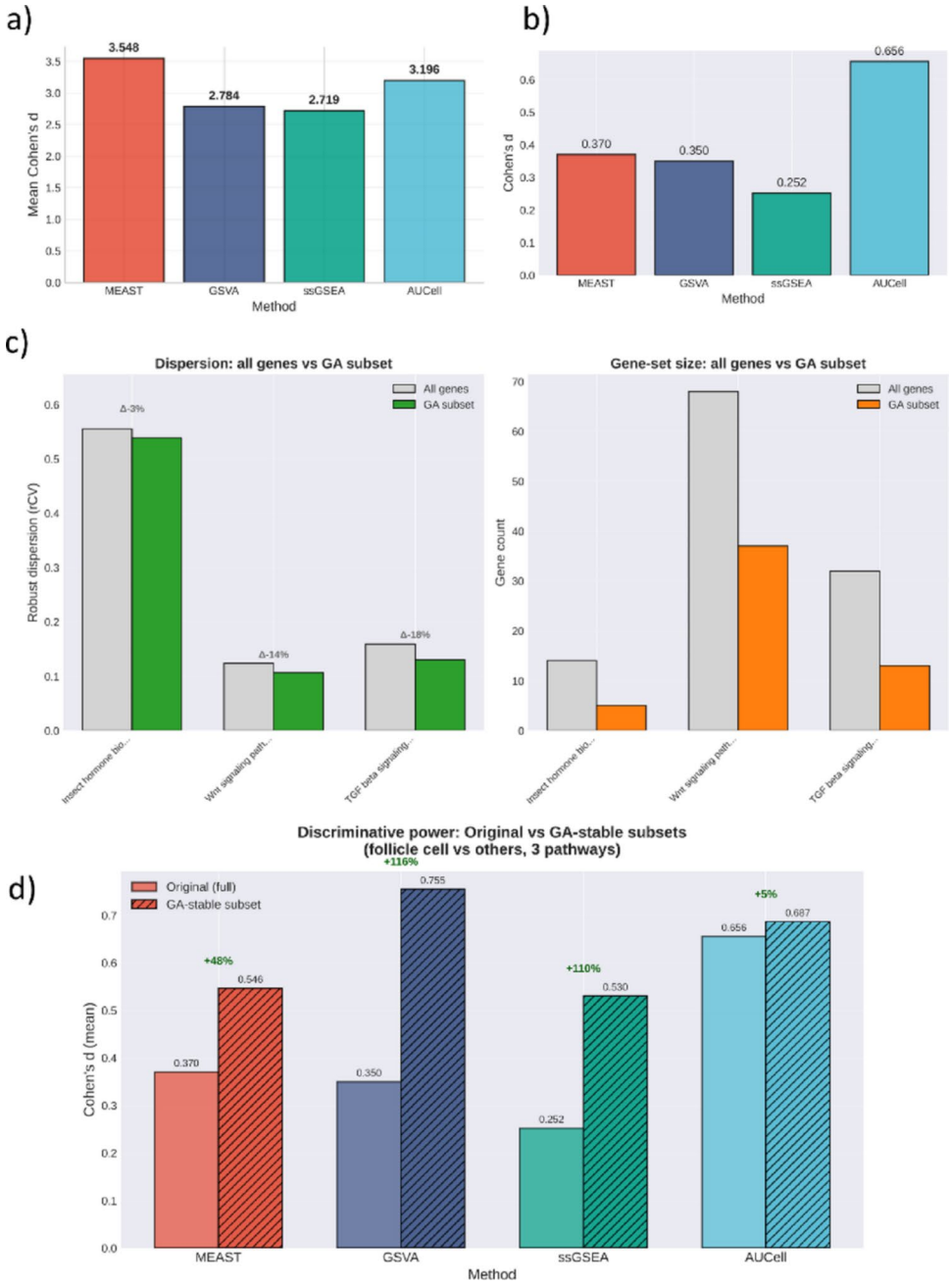

Fig. 8. Benchmarking MEAST activity scoring and evaluation of GA-based stability. (**a**) Discriminative power of activity scoring methods on the human blood dendritic cell dataset (*Villani et al.*), measured as mean Cohen's d between target and non-target cell populations across marker gene sets. MEAST shows the strongest separation compared with GSVA, ssGSEA, and AUCell. (**b**) Mean Cohen's d for three KEGG pathways in the Drosophila ovary scRNA-seq dataset (follicle cells vs. all other cells), comparing the four activity scoring methods. (**c**) Effect of the GA stability module on gene sets in the Drosophila dataset. Left: within-target score dispersion (robust coefficient of variation, rCV) for full gene sets versus GA-stable subsets. Reduced geen sets becomes more stable. Right: corresponding reduction in gene set size for each pathway. (**d**) Improvement in discriminative power after GA selection. Mean Cohen's d across the three pathways is shown for each method using the original gene sets and the GA-stable subsets, with percentage change indicated.

from 14 to 5 genes (−64.3%), Wnt signaling pathway from 68 to 37 genes (−45.6%), and TGF beta signaling pathway from 32 to 13 genes (−59.4%). This reduction improved score consistency within follicle cells, with rCV changes of −2.9%, −13.8%, and −18.2% for the three pathways, respectively (Fig. 8C).

Furthermore, GA-selected stable subsets improved discriminative power for all tested methods. Mean Cohen's d across the three pathways (follicle cell vs. others) increased after GA selection: MEAST from 0.370 to 0.546 (+47.7%), GSVA from 0.350 to 0.755 (+115.8%), ssGSEA from 0.252 to 0.530 (+110.4%), and AUCell

from 0.656 to 0.687 (+4.8%) (Fig. 8D). These results demonstrate that the GA stability module provides practical value by identifying gene subsets that enhance both stability within target populations and separation from non-target populations, regardless of the scoring method.

Discussion

This study introduces MEAST, a computational framework combining SVD-based activity scoring with GA-driven stability assessment to refine gene sets in scRNA-seq data. MEAST addresses challenges in gene set analysis by accurately scoring activity and identifying stable gene subsets. The SVD-based approach efficiently handles high dimensionality and sparsity in scRNA-seq data[31,32] resulting from biological variability, stochastic gene expression, technical biases, and library preparation[33,34]. Results from simulated datasets show MEAST effectively captures both dynamic changes and stability in gene set activity across varying dose levels (Fig. 2A, C), highlighting the utility of eigengene weights to summarize coordinated expression patterns[35].

Our method shows high efficiency for cell type activity scoring. In the human blood scRNA-seq dataset from[25], we calculated activity levels of each gene set marker. These activity scores match original results and offer greater flexibility for analysis and visualization. MEAST distinguished closely related cell types like DC2 and DC3 (Fig. S3, Table S3) and effectively differentiated LUAD from LUSC RNA-seq data using discriminative gene sets proposed by[30].

The integration of GAs for stability assessment sets MEAST apart from methods using only statistical tests or manual thresholds. GAs handle large search spaces suitable for transcriptomics analysis by selecting genes with best coexpression patterns. Our results show GA-based assessment consistently converges on stable gene subsets.

A major focus in functional genomics is achieving a balance between including enough genes to capture a biological process and excluding weakly expressed or extraneous genes that introduce noise[10,22]. MEAST tackles this challenge by using a genetic algorithm to refine initial gene sets.

The stability curve stabilizes when balance is achieved between gene number and optimal activity score. This balance depends on GA parameters, which vary with gene set complexity and size. For instance, highly complex gene set may require more GA generations. The simulated data represents a less complex dataset, where stability converge rapidly into few active genes (Fig. 5A, B). However, in the more complex dataset[25], achieving a stable state required more than 450 generations and a mutation rate of 0.001. In both cases, the method successfully reduced the gene sets to the most active genes. The stability method showed it can distinguish active subsets from noise genes and removed all added inactive genes. Applying the stability method resulted in 73% reduction for DC markers and 65% reduction for monocyte markers (Fig. 7C).

The GA relies on empirical hyperparameters (population size, mutation rate, number of generations) that may vary depending on gene set size and complexity. To assess robustness, we conducted multiple independent runs after reintroducing noise genes into previously optimized sets. In the SPARSim simulation, adding random noise genes to optimized sets and repeating the GA five times demonstrated consistent noise removal within a median of 5–6 generations across runs, with final fitness scores matching the original optimization. Similarly, in the blood scRNA-seq dataset, adding 25 random genes to the optimized pDC subset showed rapid noise exclusion by generation 6, restoring the original fitness score. Due to the stochastic nature of genetic algorithms and the inherent biological variability in gene expression data, the exact composition of optimized gene subsets may vary slightly between independent runs, particularly when multiple genes exhibit similar activity levels. However, our robustness tests demonstrate that the GA consistently converges to subsets with equivalent fitness scores and successfully excludes noise genes across repeated runs.

Analysis of convergence curves provides practical guidance for parameter selection: in the blood dataset example, the fitness curve stabilized around generation around 450, indicating that near 500 generations are sufficient for this configuration. We recommend monitoring convergence curves to ensure adequate generations for stable optimization, particularly for larger or more complex gene sets.

A limitation of our SPARSim simulation is that it primarily models differential expression with added noise but does not explicitly enforce strong coexpression structure among genes within the gene set. While dose-dependent differential expression can introduce some correlation, genes within a single condition remain largely independent. This differs from real biological gene sets, where coordinated expression patterns are common. Future work can incorporate simulations with explicit coexpression structure.

Our comparative analysis demonstrates that MEAST activity scoring performs competitively with established methods while offering a built-in scoring approach optimized for integration with the GA stability module. This integrated design ensures consistency between activity assessment and gene selection, avoiding potential discrepancies that might arise from combining incompatible methods. The GA-based gene refinement improves performance across all scoring methods tested. By removing weakly active genes, the GA module enhances signal-to-noise ratio regardless of downstream scoring approach. Our results show that the refined gene sets helps activity scoring methods to better separation of cells subtypes for example the method GSVA. These results indicate that the GA stability assessment is a gene-set refinement step that can be combined with different downstream scoring methods. MEAST is likely to be advantageous when the gene set contains a coherent co-expression structure that is well represented by a dominant eigengene and when activity scoring is used together with stability selection in a single framework. By contrast, GSVA or AUCell may be preferable when rank-based or pathway-level enrichment provides better separation in a given dataset.

Computationally, MEAST executes scoring functions within seconds and scales efficiently in high-performance environments (Tables 1 and 2), using little memory while enabling analysis of thousands of cells.

MEAST applications extend beyond this study, including identifying core gene sets for biological processes or diseases and analyzing other high-throughput data. The modular design allows integration with network inference algorithms or machine learning models. Future developments could incorporate prior biological

knowledge to improve interpretability. MEAST represents a significant advance in gene set analysis for scRNA-seq data.

Data availability

This study is based exclusively on the computational analysis of publicly available data. No primary data were collected from human participants, and no new experiments were performed on human or animal subjects. The single-cell datasets analyzed in this study were obtained from the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP43). LUAD and LUSC datasets were downloaded from The Cancer Genome Atlas Program (TCGA). Sample identifiers for all TCGA datasets used in this analysis are available in the data directory of our GitHub repository.

Code availability

The MEAST package and analysis code are available at https://gitfront.io/r/fabou/fvXUA4KYpdVR/MEAST/.

Received: 22 August 2025; Accepted: 6 April 2026
Published online: 21 April 2026

References

1. De Las, F. et al. Pathway-based genome-wide association analysis of coronary heart disease identifies biologically important gene sets. *Eur. J. Hum. Genet.* **20**, 1168–1173 (2012).
2. Pers, T. H. Gene set analysis for interpreting genetic studies. *Hum. Mol. Genet.* **25**, R133–R140 (2016).
3. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A.* **102**, 15545–15550 (2005).
4. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
5. Schlitt, T. From gene networks to gene function. *Genome Res.* **13**, 2568–2576 (2003).
6. Sprinzak, E. et al. Detecting coordinated regulation of multi-protein complexes using logic analysis of gene expression. *BMC Syst. Biol.* **3**, 115 (2009).
7. Zinani, O. Q. H., Keseroğlu, K. & Özbudak, E. M. Regulatory mechanisms ensuring coordinated expression of functionally related genes. *Trends Genet.* **38**, 73–81 (2022).
8. Dong, Z. et al. Genomic and single-cell analyses reveal genetic signatures of swimming pattern and diapause strategy in jellyfish. *Nat. Commun.* **15**, 5936 (2024).
9. Monaco, G. et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell. Rep.* **26**, 1627–1640e7 (2019).
10. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell. Syst.* **1**, 417–425 (2015).
11. Rosati, D. et al. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: a review. *Comput. Struct. Biotechnol. J.* **23**, 1154–1168 (2024).
12. Yuan, H. et al. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.* **47**, D900–D908 (2019).
13. Van De Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).
14. Wang, S. et al. Regulon active landscape reveals cell development and functional state changes of human primary osteoblasts in vivo. *Hum. Genomics*. **17**, 11 (2023).
15. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinform.* **14**, 7 (2013).
16. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
17. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*. **14**, 1083–1086 (2017).
18. Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods*. **13**, 241–244 (2016).
19. Schubert, M. et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20 (2018).
20. Vatanpour et al. Gene set analysis and reduction for a continuous phenotype: Identifying markers of birth weight variation based on embryonic stem cells and immunologic signatures. *Comput. Biol. Med.* **113**, 103389 (2019).
21. Maleki et al. Gene set analysis: challenges, opportunities, and future research. *Front. Genet.* **11**, 654 (2020).
22. Lin et al. Probabilistic prioritization of candidate pathway association with pathway score. *BMC Bioinform.* **19**, 391 (2018).
23. Yang, T. Y. A GS-CORE algorithm for performing a reduction test on multiple gene sets and their core genes. *Comput. Stat.* **30**, 29–41 (2015).
24. Lippmann, C., Ultsch, A. & Lötsch, J. Computational functional genomics-based reduction of disease-related gene sets to their key components. *Bioinformatics* **35**, 2362–2370 (2019).
25. Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
26. Tarhan, L. et al. Single Cell Portal: an interactive home for single-cell genomics data. *Preprint at.* https://doi.org/10.1101/2023.07.13.548886 (2023).
27. Shi, J. et al. A progressive somatic cell niche regulates germline cyst differentiation in the drosophila ovary. *Curr. Biol.* **31** (4), 840–852e5 (2021). Epub 2020 Dec 18. PMID: 33340458.
28. Baruzzo, G., Patuzzi, I. & Di Camillo, B. SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics* **36**, 1468–1475 (2020).
29. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
30. Chen, J. W. & Dhahbi, J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Sci. Rep.* **11**, 13323 (2021).
31. Gogolewski, K., Sykulski, M., Chung, N. C. & Gambin, A. Truncated robust principal component analysis and noise reduction for single cell RNA-seq data. *Bioinf. Res. Appl.* **10847** 335–346 (2018).
32. Tsuyuzaki et al. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biol.* **21**, 9 (2020).
33. Zhang, C. et al. Development and validation of a metastasis-associated prognostic signature based on single-cell RNA-seq in clear cell renal cell carcinoma. *Aging* **11**, 10183–10202 (2019).
34. Ji, H. et al. Decoding the cell atlas and inflammatory features of human intracranial aneurysm wall by single-cell RNA sequencing. *JAHA* **13**, e032456 (2024).
35. Kluger et al. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.* **13**, 703–716 (2003).

Acknowledgements
The authors are grateful to the Université Côte d'Azur's Center for High-Performance Computing (OPAL infrastructure) for providing resources and support. This work was also supported by BPI France through the i-Demo program.

Author contributions
F.A.C. and C.P. jointly conceived the study and designed the research framework. F.A.C. collected all data, developed the computational methodology, implemented the software, and performed the analyses. Both authors collaborated on result interpretation and discussions. F.A.C. wrote the manuscript with review input from C.P.

Funding
This work was supported by the French government through the France 2030 investment plan managed by the National Research Agency (ANR), as part of the Initiative of Excellence Université Côte d'Azur under reference number ANR-15-IDEX-01.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-48119-9.

Correspondence and requests for materials should be addressed to F.A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026